Developmental Science 7:2 (2004), pp 133–140

Autoassociator networks: insights into infant cognition

Sylvain Sirois

Department of Psychology, University of Manchester, UK

Abstract

This paper presents autoassociator neural networks. A first section reviews the architecture of these models, common learning rules, and presents sample simulations to illustrate their abilities. In a second section, the ability of these models to account for learning phenomena such as habituation is reviewed. The contribution of these networks to discussions about infant cognition is highlighted. A new, modular approach is presented in a third section. In the discussion, a role for these learning models in a broader developmental framework is proposed.

Introduction

The publication of the PDP books in 1986 is certainly a landmark event, one that introduced important new ideas to psychology. Neural networks, especially with the introduction of the backpropagation learning rule for multilayered networks, generated significant interest and controversy. The impact of connectionism on developmental psychology in particular has been highlighted in recent publications (e.g. Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996; Quinlan, 2003). But neural networks had been around since the 1940s; they essentially went out of the mainstream research by the late 1960s. There were still researchers working on neural networks at that time, though. And one class of networks they worked with are called autoassociators (Anderson, Silverstein, Ritz & Jones, 1977; Kohonen, 1977).

In this paper, I introduce autoassociator networks, beginning with an overview of their architecture and how they work. Simple simulations highlight how they process information. I then review recent work that models infancy phenomena with autoassociator networks, and how these simulations can inform developmental psychologists. A new approach to habituation, using autoassociators in a modular framework, is presented. In the discussion, I suggest how such learning processes instruct the broader issue of accounting for cognitive change in a developmental framework.

Autoassociators: the nuts and bolts

Autoassociator networks consist of a single bank of interconnected units (see Figure 1). A stimulus is presented to these units, activating them to various degrees, and the resulting activations are circulated within the network. A stimulus is normally represented as a vector of numerical values that correspond to various features. The units are typically linear integrators, meaning that they sum the input they receive in order to compute their activations. In most autoassociators, each unit sends its activation to every other unit through weighted connections. These weights, represented by real numbers, determine the magnitude of stimulation or inhibition. Usually, the weight between a unit and itself is deleted, such that the activation of a unit is solely a function of the activations of other units (and stimulus, when applicable).

Information processing and learning

Information processing in autoassociator networks usually proceeds as follows. A stimulus is presented to the network, whereby each unit takes as activation value the value of the specific feature it is presented with. The external stimulus may remain *clamped* on, in which case it remains part of the units' input on subsequent cycles. The stimulus may fade over time, in which case a progressively smaller proportion of the features' values are part of the input on subsequent cycles. On processing

Address for correspondence: Sylvain Sirois, Department of Psychology, The University of Manchester, Oxford Road, Manchester M13 9PL, UK; e-mail: sylvain.sirois@man.ac.uk

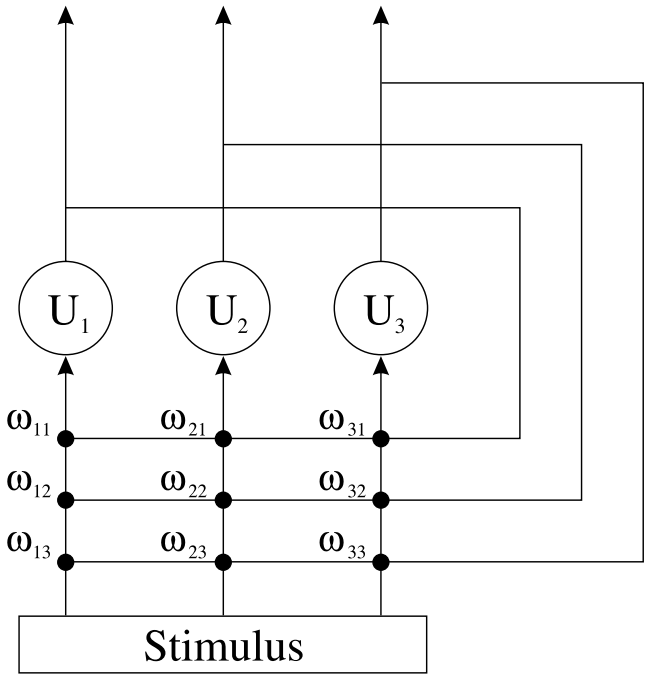

Figure 1 *Depiction of an autoassociator, consisting of three units (U_1 to U_3). External stimuli are applied to units, which cycle their activations, stimulating or inhibiting one another. This is achieved by sending activations through weighted connections, labelled* w_{ij}, *where* i *is the receiving unit, and* j *the receiving unit.*

cycles where a stimulus is no longer presented, the only source of input to a unit is the activity of units in the network. This typically goes on for a fixed number of processing cycles, or until most or all units have settled on stable activation values.[1]

Therefore, at any time following the presentation of a stimulus, the net input of a unit may be computed as

$$net_i = \sum (w_{ij} a_j) \tag{1}$$

where net_i is the net input received by unit i, w_{ij} is the weight between sending unit j and receiving unit i, and a_j is the current activation of unit j. When a stimulus S remains clamped on, the value of feature S_i (or a proportion of it, for fading stimuli) is added to net_i. The activation of a unit is computed from its net input. For linear units, the activation is equal to the net input. For asigmoid units, such as used later in this paper, the net input is passed through the function

[1] It should be noted that there is no direct and unambiguous way to equate processing cycles with real time. Such mappings, when necessary, are typically arbitrary but should be minimally consistent within the scope of a simulation (Sirois, Buckingham & Shultz, 2000).

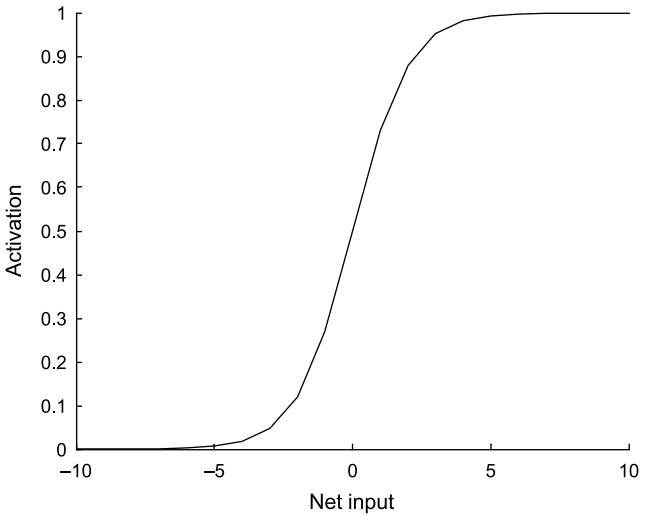

Figure 2 *Activation values as a function of net input for the asigmoid function.*

$$a_i = \frac{1}{1 + \exp(-net_i)} \tag{2}$$

where a_i is the activation of unit i, and net_i is the net input to unit i. This function produces s-shaped activation values constrained between 0 and 1, such as shown in Figure 2.

The most common use of autoassociator networks is learning to reproduce stimuli. A reason for this is that these networks are especially apt at reconstructing known stimuli from partial or noisy input. Common learning rules capitalize on correlations between units' activations. One such rule is the simple Hebbian rule, expressed as

$$\Delta w_{ij} = \lambda a_i a_j \tag{3}$$

where Δw_{ij} is the amount by which to change the weight between receiving unit i and sending unit j, λ is the learning rate, a_i the activation of receiving unit i, and a_j the activation of sending unit j. This equation ensures that units with shared activity are linked by larger weights than units with little covariance. The learning rate is usually a value between 0 and 1 that ensures that weights do not change too much, too quickly. This is important when a network has to learn many patterns. A large learning rate would bias weights towards the first few stimuli encountered.

A problem with the Hebbian rule is that weights can grow unconstrained, leading to an overtrained network. An alternative is to use the delta rule, or Widrow-Hoff rule, which ensures that weights stay within acceptable bounds. The rule is expressed as

$$\Delta w_{ij} = \lambda (s_i - a_i) a_j \tag{4}$$

Figure 3 *The first row of this figure shows the stimuli used to train autoassociator networks (and the rightmost bar shows the range of pixel values). Test images are shown on the second row. These are modifications of the training images, with the exception of the rightmost image, used to stimulate units uniformly and produce a prototypical pattern of activations. The third and fourth rows show test activations for an autoassociator and a novelty filter, respectively.*

where Δw_{ij} is the amount by which to change the weight between receiving unit i and sending unit j, λ is the learning rate, s_i the stimulus value at feature i, a_i the activation of receiving unit i, and a_j the activation of sending unit j. This rule ensures that weight changes are progressively smaller as activations approach target values.

An autoassociator example

An example may prove helpful at this stage. Although not a model of face perception, the ability of the autoassociator to learn a set of pictures of faces and 'recognize' them when distorted will provide readers with a clear illustration of what these networks can do. The first row of images in Figure 3 shows six faces with which an autoassociator was trained using the delta rule. Every pixel position in this set of images was associated with a specific unit in an autoassociator network, and specific pixel values in given images served as input to the network. The low-resolution images were arrays of 51 by 70 pixels, and thus required 3570 units to represent, which imply 12,744,900 weights. The second row of images contains seven images with which the network was tested after training. The first six images are modifications of the training images. A rectangular mask was applied to the first training image, whereas different filters were applied to the remaining five images, resulting in various amounts of noise and distortion. The seventh test image is a grey field, with pixels set at the middle value of 0.5. The purpose of this image is to equally activate all units in the network, after which the network is left to settle on a pattern of activations that illustrates the prototypical representation it has learned.

The third row in Figure 3 shows the activations recorded in an autoassociator on the seven test patterns after eight epochs of learning with the six training images (an epoch consists of one presentation of all training patterns, and so the network had 'seen' 48 faces before testing). As can be seen, the mask has been removed from test image 1, and the various distortions on test images 2–4 were smoothed out. Test image 5 was probably too distorted, and the reconstructed face does not retain the unique features of the original. It is, effectively, averaged from all known faces in the reconstruction process. Test image 6 was bright and saturated, and thus the reconstructed image, while retaining the unique features of the original face, is somewhat noisy. Finally, when tested with a grey image, the network produces an average face, one that reflects the differences of the various training faces but which remains quite face-like.

Overall, the autoassociator does a reasonable job of representing the six training faces in this toy demonstration. Although what it effectively learns is an average face, it preserves to an extent some of the unique features of the training faces. When these are presented with moderate distortion, the network can *clean up* the stimuli, acting as a filter.

The idea of the autoassociator as a filter is explicit in a variant called the novelty filter (Kohonen, 1988). This network uses a learning rule expressed as

$$\Delta w_{ij} = \lambda(floor - a_i)a_j \quad (5)$$

where Δw_{ij} is the amount by which to change the weight between receiving unit i and sending unit j, λ is the learning rate, *floor* is the minimum value of the activation function, a_i the activation of receiving unit i, and a_j the activation of sending unit j. Using this rule, networks learn to inhibit the activations of correlated units. If units a_i and a_j are both active, the weights between them are decreased as a function of how far the receiving unit is away from its floor activation. Weight changes thus tend to be smaller over time. Over training, units in a novelty filter should progressively stop responding when shown known stimuli. The fourth row of Figure 3 shows activations in a novelty filter for test stimuli following training on the six faces used previously. Darker pixels represent inhibited units, whereas brighter pixels represent active units (i.e. responding to some novelty).

The mask from the first test image has an interesting effect on the novelty filter. Although the mask is clearly active outside the face area (and thus novel), the face region of the mask is more inhibited than the rest of the face. This is because the filter was trained on pixels and not some feature-based translation of the images. The pixels most activated in the set of training images are in the face area, and thus this area becomes more inhibited. With the first test image, the uniform band provides ample external stimulation to neighbouring units that typically covary, resulting in inhibition in the face region (where units are usually active) and activation in the periphery (where units are usually inactive). In the remaining five test faces, it can be seen that the network responds most (bright pixels) to specific attributes of the individual faces (such as hairlines, eyes and mouths), as well as unusual patterns of activations (the various distortions). Finally, the response of the novelty filter to a grey test image is a negative face. The regions less inhibited, and thus more 'interesting' within the face are the eyes, mouth and nose outline. This is because they are regions associated with darker pixels, and not because these are perceived as 'important' features. Nevertheless, when a face is presented to the novelty filter, what *pops out* is what we think of as important features, which also prove important early in infancy (Johnson & Morton, 1991).

Autoassociators and development

Autoassociators are not developmental models, because the representational power of a network never grows. Networks represent stimuli within an activation space defined by the number of units and the activation function. This space does not change as a function of experience. What experience accounts for is where in that space a network will end up when presented a given stimulus. This is learning (Quartz, 1993; Sirois & Shultz, 2003).

So what may autoassociator networks do for developmentalists? They can actually provide useful insights when claims about cognitive abilities might be confounded with simpler learning processes. An example is infant cognition using habituation. Using tasks derived from the habituation tradition, researchers have suggested a host of complex abilities in young infants, including object permanence (Baillargeon, 1987), knowledge of physics (Spelke, Breinlinger, Macomber & Jacobson, 1992), number (Wynn, 1992) and language (Marcus, Vijayan, Bandi Rao & Vishton, 1999). Claims of complex cognitive abilities in infants were criticized, however, on the grounds that these failed to consider simpler perceptual learning mechanisms or various methodological issues (Bogartz, Shinskey & Speaker, 1997; Haith, 1998). A working example of perceptual learning might therefore help resolve the debate.

Habituation as autoassociation

Kohonen (1988) suggested that the novelty filter could model habituation. Such networks progressively stop responding to a training set, and show renewed interest in novel patterns. Although it has been used in some robotics applications (e.g. Marsland, Nehmzow & Shapiro, 2000), I have failed to find an application to infancy data. However, a model of habituation based on the autoassociator framework was recently published (Sirois *et al.*, 2000). An overview of the original infancy study, its conclusions and the subsequent arguments made from the modelling will highlight the usefulness of the approach to cognitive development.

Researchers familiarized 7-month-old infants to a series of three syllables that followed a systematic pattern, in order to assess whether infants have the ability to learn grammar-type rules (Marcus *et al.*, 1999). With an 'aba' pattern, the first and third syllable were identical (e.g. 'li-na-li', 'ga-ti-ga'). With an 'abb' pattern, the second and third syllables were the same (e.g. 'ga-ti-ti', 'li-na-na'). After familiarization with one pattern, infants were alternately presented with new syllables that followed either of the pattern structures. Infants familiarized with 'aba' patterns responded more to 'abb' test patterns and vice versa. The effect was robust to replications that instantiated various controls.

The authors argued that the use of new syllables for test patterns implied that alternative statistical learning interpretations (e.g. Saffran, Aslin & Newport, 1996)

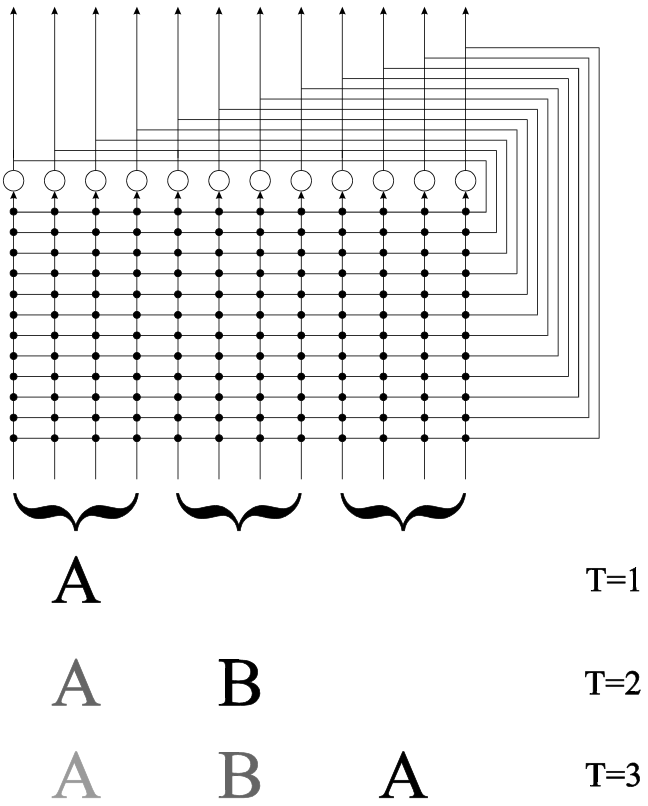

Figure 4 *Network used by Sirois* et al. *(2000). Syllables were introduced sequentially and faded over processing cycles.*

could not account for the data, nor could systems that represent strings as made up of tokens. Moreover, the authors argued that neural networks would fail to reproduce infants' behaviour, and reported unsuccessful simulations using simple recurrent networks (SRNs, a class of feedforward, multilayered neural network). The authors thus argued that infants had a pre-linguistic ability to extract abstract rules from speech streams, and could thus appreciate violations of simple grammatical rules. Such an early ability would provide crucial support to influential theories of language acquisition (Pinker, 1999).

Paradoxically, the claim that neural networks would fail has made this study one of the most successfully modelled phenomena (Shultz & Bale, 2001). Of the 10 or so models that have appeared, only one used autoassociator networks (Sirois *et al.*, 2000). The networks, depicted in Figure 4, consisted of 12 interconnected units. The three 'syllables' that comprised a 'sentence' were presented sequentially, each to a unique bank of four units. These syllables were coded as simple patterns of binary values (e.g. [1 0 1 1]). This approach essentially trades time for space. A fading encoding scheme was used, such that input present at time step 1 was only a proportion of itself at time step 2, and so forth. Networks were trained on one 'grammar' (e.g. 'aba') for a fixed number of trials, and then tested on examples of two grammars (e.g. 'aba' and 'abb'), using new syllables. The dependent measure on test trials was the number of cycles networks required to settle (i.e. activations changed by less than some criterion) on test patterns. The assumption is that such processing time is analogous to what sustains overt interest in infants.

The networks reproduced the pattern of behaviour observed with infants. Networks required significantly more time to settle on patterns that violated the structure of the training set. As other simulations of this task have shown, neural networks can indeed capture the behaviour of infants on this learning task. The only requirement to do so, in the case of the autoassociator simulations, is the ability to capture correlations between features of sequential stimuli.

This does not show that infants do not use abstract rules. However, it does show that these rules are not necessary, because simple statistical learning can account for the data equally well, and more parsimoniously. Autoassociator networks can thus provide insights into issues about infant cognition, which bear general implications for developmental psychologists.

Networked networks

Autoassociators were used to model habituation because they implemented some key features of the task that eluded alternative models, while avoiding some of their computational limitations (Sirois *et al.*, 2000). However, a recent review of models of habituation identified seven key features of the phenomena that models should strive to accommodate (Sirois & Mareschal, 2002a). There are five crucial behavioural features:

1. Temporal unfolding of behaviours
2. Exponential decrease of responses
3. A shift from familiarity to novelty preference over trials
4. Habituation to repeated testing
5. The ability to discriminate habituated items

Although models of habituation need not be models of neural circuits, they should be consistent with basic neural features associated with habituation:

1. Hippocampal selective inhibition
2. Subcortical-cortical interactions

Autoassociators only capture behavioural features 1, 2, 4 and 5, whereas the novelty filter captures behavioural features 1, 2 and 4, as well as neural feature 1. Although they can model habituation data, they cannot

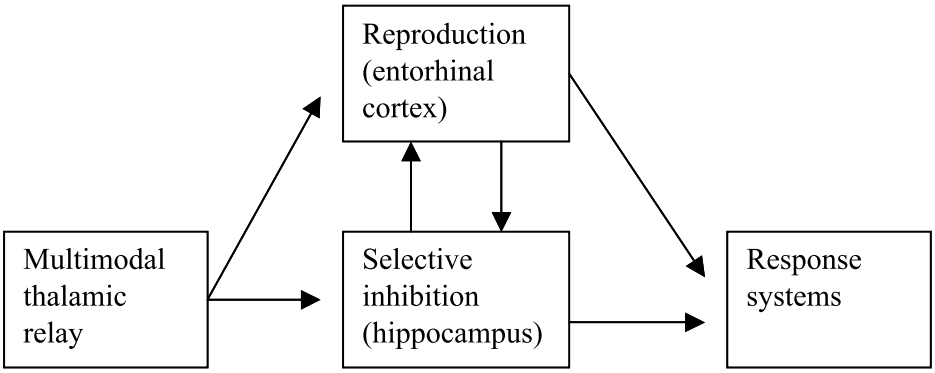

Figure 5 *Schematic depiction of neural functions that support habituation, and their relations.*

be considered appropriate models of habituation *per se*. Currently, no model of habituation successfully captures all key behavioural and neural features.

In order to address previous limitations, a new, general-purpose model of habituation was proposed (Sirois & Mareschal, 2002b). This new model, which may be monikered HAB (for Habituation, Autoassociation and Brain), was built by considering the functional organization of neural circuits involved in habituation, shown schematically in Figure 5. In habituation, the hippocampus plays a crucial role of selective inhibition of stimulus features common to the habituation set (Vinogradova, 1975). The hippocampus also interacts with cortical structures (especially the entorhinal cortex), affecting habituation behaviour (Nelson, 2002). Indeed, whereas the hippocampus is involved in fast and transient computations, cortical areas are involved in long-term information encoding (necessary to account for long-term effects of habituation). Moreover, cortical structures provide the gateway for effects of prior experience onto habituation.

In the HAB model, a novelty filter network implements hippocampal selective inhibition, whereas cortical reproduction is realized with a standard autoassociator. Weights between the sub-networks allow each to alter the input of the other. The cortical network attempts to reproduce known input, and thus amplifies corresponding input features to the hippocampal network. The latter tries to inhibit known input, and thus weakens known features of the cortical inputs. The output of both sub-networks is combined to produce the system's output. This architecture, by design, satisfies the key neural constraints of habituation.

It turns out that this neurally constrained model produces the crucial behavioural markers of habituation (Sirois & Mareschal, 2002b): behaviour unfolds over time naturally in autoassociators, activations decrease exponentially over learning, there is a shift from familiarity to novelty preference in network activations during training, habituation to repeated testing, and the ability to discriminate between familiar stimuli (due to active representations in the cortical network). It is most promising that the shift from familiarity to novelty emerges from a model consistent with other behavioural markers as well as with neural features of habituation. This familiarity-novelty shift is increasingly recognized as a crucial feature in discussions of infant habituation and cognition (Cohen & Marks, 2002).

Discussion

Autoassociator networks are a class of neural networks with minimal assumptions about internal representations. They have, however, many degrees of freedom from connection weights, which increase exponentially with every unit added to a network (i.e. the total number of weights is the square of the number of units, minus the number of units if self-connections are prevented). This allows networks to learn a relatively large number of unique stimuli, but it also makes examining internal representations nearly intractable, especially as the number of units grows large. However, learning is based on local computations and resulting weights are straightforward in interpretation (i.e. large weights imply large correlations between features).

This paper has shown how such simple learning devices can offer insights for developmental research. These networks can reproduce behaviour deemed to reflect advanced cognitive skills, and thus suggest caution when interpreting data. As the HAB model illustrates, autoassociator networks that implement specific functions in a neurally constrained modular framework can capture a wider array of behaviours. Moreover, this

framework is designed to bridge the gap between neural and behavioural sciences. Combined with the relative parsimony of the model, this consistency across levels of interpretation gives the model unparalleled appeal.

Assuming the HAB model proved an adequate model of early infant learning, the obvious question of the developmentalist would be: what next? If infants were shown not to hold the various abilities that some researchers have ascribed to them, older children and adults certainly do.

As discussed earlier, autoassociator networks are not developmental models. However, HAB is a network of networks and, although hardwired, it illustrates one promising avenue of research. Combining two autoassociators in a single system allowed capturing behaviours that eluded the individual component networks. This is the very idea of development: to recruit previously independent processes into a more powerful structure that qualitatively changes what and how information is processed. This notion was at the core of Piaget's abstraction and Karmiloff-Smith's representational redescription (Sirois & Shultz, 2003).

Modelling data with single networks that start from scratch may be sufficient to achieve the goals of a simulation. This has been the most frequent approach so far. But thanks partly to faster and cheaper computers, developmental researchers can begin to look at the broader picture of cognitive change, one of embedded systems that make use of both prior knowledge and structural changes in order to progress from simple learning to complex thinking.

There is already some promising work in this area. Shultz and Rivest (2001) have adapted the Cascade-correlation neural network model in order to recruit old networks into new networks. This makes efficient use of prior knowledge in a truly developmental framework (Sirois & Shultz, 2003).

The publication of the PDP books in 1986 marked the start of a new era for psychology. And the usefulness of neural networks to developmental psychology was possibly best captured in *Rethinking innateness* (Elman *et al.*, 1996). It may now be the time to start thinking of development as networks that become part of broader networks, which become part of ever-broader networks. But at the onset, networks could be quite simple.

References

Anderson, J.A., Silverstein, J.W., Ritz, S.A., & Jones, R.S. (1977). Distinctive features, categorical perception, and probability learning: some applications of a neural model. *Psychological Review*, **84**, 413–451.

Baillargeon, R. (1987). Object permanence in 3.5- and 4.5-month-old infants. *Developmental Psychology*, **23**, 655–664.

Bogartz, R.S., Shinskey, J.L., & Speaker, C.J. (1997). Interpreting infant looking: the event set X event set design. *Developmental Psychology*, **33**, 408–422.

Cohen, L.B., & Marks, K.S. (2002). How infants process addition and subtraction events. *Developmental Science*, **5**, 186–201.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Haith, M.M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior & Development*, **21**, 167–179.

Johnson, M.H., & Morton, J. (1991). *Biology and cognitive development: The case of face perception*. Oxford: Blackwell.

Kohonen, T. (1977). *Associative memory: A system theoretical approach*. New York: Springer.

Kohonen, T. (1988). *Self-organization and Associative memory*, 2nd edition. New York: Springer-Verlag.

Marcus, G.F., Vijayan, S., Bandi Rao, S., & Vishton, P.M. (1999). Rule learning by seven-month-old infants. *Science*, **283**, 77–80.

Marsland, S., Nehmzow, U., & Shapiro, J. (2000). Novelty detection on a mobile robot using habituation. *From Animals to Animats: Proceedings of the 6th International Conference on Simulation of Adaptive Behaviour (SAB'00)* (pp. 189–198). Cambridge, MA: MIT Press.

Nelson, C.A. (2002). The ontogeny of human memory: a cognitive neuroscience perspective. In M.H. Johnson, Y. Munakata & R.O. Gilmore (Eds.), *Brain development and cognition: A reader* (2nd edn., pp. 151–178). Malden, MA: Blackwell.

Pinker, S. (1999). Out of the minds of babes. *Science*, **283**, 40–41.

Quartz, S.R. (1993). Neural networks, nativism, and the plausibility of constructivism. *Cognition*, **48**, 223–242.

Quinlan, P. (Ed.) (2003). *Connectionist models of development*. Hove, England: Psychology Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Shultz, T.R., & Bale, A.C. (2001). Neural network simulation of infant familiarization to artificial sentences: rule-like behavior without explicit rules and variables. *Infancy*, **2**, 501–536.

Shultz, T.R., & Rivest, F. (2001). Knowledge-based cascade-correlation: using knowledge to speed learning. *Connection Science*, **13**, 43–72.

Sirois, S., Buckingham, D., & Shultz, T.R. (2000). Artificial grammar learning by infants: an auto-associator perspective. *Developmental Science*, **4**, 442–456.

Sirois, S., & Mareschal, D. (2002a). Computational approaches to infant habituation. *Trends in Cognitive Sciences*, **6**, 293–298.

Sirois, S., & Mareschal, D. (2002b). Infant habituation: a review of current computational models and a new proposal. In J.A. Bullinaria & W. Lowe (Eds.), *Proceedings of the*

Seventh Neural Computation and Psychology Workshop: Connectionist Models of Cognition and Perception (pp. 90–103). Singapore: World Scientific.

Sirois, S., & Shultz, T.R. (2003). A connectionist perspective on Piagetian development. In P. Quinlan (Ed.), *Connectionist models of development* (pp. 13–41). Hove, England: Psychology Press.

Spelke, E.S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, **99**, 605–632.

Vinogradova, O.S. (1975). The hippocampus and the orienting reflex. In E.N. Sokolov & O.S. Vinogradova (Eds.), *Neuronal mechanisms of the orienting reflex* (pp. 128–154). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, **358**, 749–750.